AI SAFETY GOVERNANCE FRAMEWORK

National Technical Committee 260 on Cybersecurity of SAC
2024.9

Content

AI Safety Governance Framework (V1.0)

Artificial Intelligence (AI), a new area of human development, presents significant opportunities to the world while posing various risks and challenges. Upholding a people-centered approach and adhering to the principle of developing AI for good, this framework has been formulated to implement the Global AI Governance Initiative and promote consensus and coordinated efforts on AI safety governance among governments, international organizations, companies, research institutes, civil organizations, and individuals, aiming to effectively prevent and defuse AI safety risks.

1. Principles for AI safety governance

-Commit to a vision of common, comprehensive, cooperative, and sustainable security while putting equal emphasis on development and security

-Prioritize the innovative development of AI

-Take effectively preventing and defusing AI safety risks as the starting point and ultimate goal

-Establish governance mechanisms that engage all stakeholders, integrate technology and management, and ensure coordinated efforts and collaboration among them

-Ensure that all parties involved fully shoulder their responsibilities for AI safety

-Create a whole-process, all-element governance chain

-Foster a safe, reliable, equitable, and transparent AI for the technical research, development, and application

-Promote the healthy development and regulated application of AI

-Effectively safeguard national sovereignty, security and development interests

-Protect the legitimate rights and interests of citizens, legal persons and other organizations

-Guarantee that AI technology benefits humanity

1.1 Be inclusive and prudent to ensure safety

We encourage development and innovation and take an inclusive approach to AI research, development, and application. We make every effort to ensure AI safety, and will take timely measures to address any risks that threaten national security, harm the public interest, or infringe upon the legitimate rights and interests of individuals.

1.2 Identify risks with agile governance

By closely tracking trends in AI research, development, and application, we identify AI safety risks from two perspectives: the technology itself and its application. We propose tailored preventive measures to mitigate these risks. We follow the evolution of safety risks, swiftly adjusting our governance measures as needed. We are committed to improving the governance mechanisms and methods while promptly responding to issues warranting government oversight.

1.3 Integrate technology and management for coordinated response

We adopt a comprehensive safety governance approach that integrates technology and management to prevent and address various safety risks throughout the entire process of AI research, development, and application. Within the AI research, development, and application chain, it is essential to ensure that all relevant parties, including model and algorithm researchers and developers, service providers, and users, assume their respective responsibilities for AI safety. This approach well leverages the roles of governance mechanisms involving government oversight, industry self-regulation, and public scrutiny.

1.4 Promote openness and cooperation for joint governance and shared benefits

We promote international cooperation on AI safety governance, with the best practices shared worldwide. We advocate establishing open platforms and advance efforts to build broad consensus on a global AI governance system through dialogue and cooperation across various disciplines, fields, regions, and nations.

2. Framework for AI safety governance

Based on the notion of risk management, this framework outlines control measures to address different types of AI safety risks through technological and managerial strategies. As AI research, development, and application rapidly evolves, leading to changes in the forms, impacts, and our perception of safety risks, it is necessary to continuously update control measures, and invite all stakeholders to refine the governance framework.

2.1 Safety and security risks

By examining the characteristics of AI technology and its application scenarios across various industries and fields, we pinpoint safety and security risks and potential dangers that are inherently linked to the technology itself and its application.

2.2 Technical countermeasures

Regarding models and algorithms, training data, computing facilities, products and services, and application scenarios, we propose targeted technical measures to improve the safety, fairness, reliability, and robustness of AI products and applications. These measures include secure software development, data quality improvement, construction and operations security enhancement, and conducting evaluation, monitoring, and reinforcement activities.

2.3 Comprehensive governance measures

In accordance with the principle of coordinated efforts and joint governance, we clarify the measures that all stakeholders, including technology research institutions, product and service providers, users, government agencies, industry associations, and social organizations, should take to identify, prevent, and respond to AI safety risks.

2.4 Safety guidelines for AI development and application

We propose several safety guidelines for AI model and algorithm developers, AI service providers, users in key areas, and general users, to develop and apply AI technology.

3. Classification of AI safety risks

Safety risks exist at every stage throughout the AI chain, from system design to research and development (R&D), training, testing, deployment, utilization, and maintenance. These risks stem from inherent technical flaws as well as misuse, abuse, and malicious use of AI.

3.1 AI's inherent safety risks

3.1.1 Risks from models and algorithms

(a) Risks of explainability

AI algorithms, represented by deep learning, have complex internal workings. Their black-box or grey-box inference process results in unpredictable and untraceable outputs, making it challenging to quickly rectify them or trace their origins for accountability should any anomalies arise.

(b) Risks of bias and discrimination

During the algorithm design and training process, personal biases may be introduced, either intentionally or unintentionally. Additionally, poor-quality datasets can lead to biased or discriminatory outcomes in the algorithm's design and outputs, including discriminatory content regarding ethnicity, religion, nationality and region.

(c) Risks of robustness

As deep neural networks are normally non-linear and large in size, AI systems are susceptible to complex and changing operational environments or malicious interference and inductions, possibly leading to various problems like reduced performance and decision-making errors.

(d) Risks of stealing and tampering

Core algorithm information, including parameters, structures, and functions, faces risks of inversion attacks, stealing, modification, and even backdoor injection, which can lead to infringement of intellectual property rights (IPR) and leakage of business secrets. It can also lead to unreliable inference, wrong decision output and even operational failures.

(e) Risks of unreliable output

Generative AI can cause hallucinations, meaning that an AI model generates untruthful or unreasonable content, but presents it as if it were a fact, leading to biased and misleading information.

(f) Risks of adversarial attack

Attackers can craft well-designed adversarial examples to subtly mislead, influence and even manipulate AI models, causing incorrect outputs and potentially leading to operational failures.

3.1.2 Risks from data

(a)Risks of illegal collection and use of data

The collection of AI training data and the interaction with users during service provision pose security risks, including collecting data without consent and improper use of data and personal information.

(b)Risks of improper content and poisoning in training data

If the training data includes illegal or harmful information like false, biased and IPR-infringing content, or lacks diversity in its sources, the output may include harmful content like illegal, malicious, or extreme information. Training data is also at risk of being poisoned from tampering, error injection, or misleading actions by attackers. This can interfere with the model's probability distribution, reducing its accuracy and reliability.

(c)Risks of unregulated training data annotation

Issues with training data annotation, such as incomplete annotation guidelines, incapable annotators, and errors in annotation, can affect the accuracy, reliability, and effectiveness of models and algorithms. Moreover, they can introduce training biases, amplify discrimination, reduce generalization abilities, and result in incorrect outputs.

(d) Risks of data leakage

In AI research, development, and applications, issues such as improper data processing, unauthorized access, malicious attacks, and deceptive interactions can lead to data and personal information leaks.

3.1.3 Risks from AI systems

(a)Risks of exploitation through defects and backdoors

The standardized API, feature libraries, toolkits used in the design, training, and verification stages of AI algorithms and models, development interfaces, and execution platforms, may contain logical flaws and vulnerabilities. These weaknesses can be exploited, and in some cases, backdoors can be intentionally embedded, posing significant risks of being triggered and used for attacks.

(b) Risks of computing infrastructure security

The computing infrastructure underpinning AI training and operations, which relies on diverse and ubiquitous computing nodes and various types of computing resources, faces risks such as malicious consumption of computing resources and cross-boundary transmission of security threats at the layer of computing infrastructure.

(c) Risks of supply chain security

The AI industry relies on a highly globalized supply chain. However, certain countries may use unilateral coercive measures, such as technology barriers and export restrictions, to create development obstacles and maliciously disrupt the global AI supply chain. This can lead to significant risks of supply disruptions for chips, software, and tools.

3.2 Safety risks in AI applications

3.2.1 Cyberspace risks

(a) Risks of information and content safety

AI-generated or synthesized content can lead to the spread of false information, discrimination and bias, privacy leakage, and infringement issues, threatening the safety of citizens' lives and property, national security, ideological security, and causing ethical risks. If users' inputs contain harmful content, the model may output illegal or damaging information without robust security mechanisms.

(b) Risks of confusing facts, misleading users, and bypassing authentication

AI systems and their outputs, if not clearly labeled, can make it difficult for users to discern whether they are interacting with AI and to identify the source of generated content. This can impede users' ability to determine the authenticity of information, leading to misjudgment and misunderstanding. Additionally, AI-generated highly realistic images, audio, and videos may circumvent existing identity verification mechanisms, such as facial recognition and voice recognition, rendering these authentication processes ineffective.

(c) Risks of information leakage due to improper usage

Staff of government agencies and enterprises, if failing to use the AI service in a regulated and proper manner, may input internal data and industrial information into the AI model, leading to leakage of work secrets, business secrets and other sensitive business data.

(d) Risks of abuse for cyberattacks

AI can be used in launching automatic cyberattacks or increasing attack efficiency, including exploring and making use of vulnerabilities, cracking passwords, generating malicious codes, sending phishing emails, network scanning, and social engineering attacks. All these lower the threshold for cyberattacks and increase the difficulty of security protection.

(e) Risks of security flaw transmission caused by model reuse

Re-engineering or fine-tuning based on foundation models is commonly used in AI applications. If security flaws occur in foundation models, it will lead to risk transmission to downstream models.

3.2.2 Real-world risks

(a)Inducing traditional economic and social security risks

AI is used in finance, energy, telecommunications, traffic, and people's livelihoods, such as self-driving and smart diagnosis and treatment. Hallucinations and erroneous decisions of models and algorithms, along with issues such as system performance degradation, interruption, and loss of control caused by improper use or external attacks, will pose security threats to users' personal safety, property, and socioeconomic security and stability.

(b) Risks of using AI in illegal and criminal activities

AI can be used in traditional illegal or criminal activities related to terrorism, violence, gambling, and drugs, such as teaching criminal techniques, concealing illicit acts, and creating tools for illegal and criminal activities.

(c) Risks of misuse of dual-use items and technologies

Due to improper use or abuse, AI can pose serious risks to national security, economic security, and public health security, such as greatly reducing the capability requirements for non-experts to design, synthesize, acquire, and use nuclear, biological, and chemical weapons and missiles; designing cyber weapons that launch network attacks on a wide range of potential targets through methods like automatic vulnerability discovering and exploiting.

3.2.3 Cognitive risks

(a) Risks of amplifying the effects of "information cocoons"

AI can be extensively utilized for customized information services, collecting user information, and analyzing types of users, their needs, intentions, preferences, habits, and even mainstream public awareness over a certain period. It can then be used to offer formulaic and tailored information and service, aggravating the effects of "information cocoons."

(b) Risks of usage in launching cognitive warfare

AI can be used to make and spread fake news, images, audio, and videos, propagate content of terrorism, extremism, and organized crimes, interfere in internal affairs of other countries, social systems, and social order, and jeopardize sovereignty of other countries. AI can shape public values and cognitive thinking with social media bots gaining discourse power and agenda-setting power in cyberspace.

3.2.4 Ethical risks

(a)Risks of exacerbating social discrimination and prejudice, and widening the intelligence divide

AI can be used to collect and analyze human behaviors, social status, economic status, and individual personalities, labeling and categorizing groups of people to treat them discriminatingly, thus causing systematical and structural social discrimination and prejudice. At the same time, the intelligence divide would be expanded among regions.

(b)Risks of challenging traditional social order

The development and application of AI may lead to tremendous changes in production tools and relations, accelerating the reconstruction of traditional industry modes, transforming traditional views on employment, fertility, and education, and bringing challenges to stable performance of traditional social order.

(c)Risks of AI becoming uncontrollable in the future

With the fast development of AI technologies, there is a risk of AI autonomously acquiring external resources, conducting self-replication, become self-aware, seeking for external power, and attempting to seize control from humans.

4. Technological measures to address risks

Responding to the above risks, AI developers, service providers, and system users should prevent risks by taking technological measures in the fields of training data, computing infrastructures, models and algorithms, product services, and application scenarios.

4.1 Addressing AI's inherent safety risks

4.1.1 Addressing risks from models and algorithms

(a) Explainability and predictability of AI should be constantly improved to provide clear explanation for the internal structure, reasoning logic, technical interfaces, and output results of AI systems, accurately reflecting the process by which AI systems produce outcomes.

(b) Secure development standards should be established and implemented in the design, R&D, deployment, and maintenance processes to eliminate as many security flaws and discrimination tendencies in models and algorithms as possible and enhance robustness.

4.1.2 Addressing risks from data

(a) Security rules on data collection and usage, and on processing personal information should be abided by in all procedures of training data and user interaction data, including data collection, storage, usage, processing, transmission, provision, publication, and deletion. This aims to fully ensure user's legitimate rights stipulated by laws and regulations, such as their rights to control, to be informed, and to choose.

(b) Protection of IPR should be strengthened to prevent infringement on IPR in stages such as selecting training data and result outputs.

(c) Training data should be strictly selected to ensure exclusion of sensitive data in high-risk fields such as nuclear, biological, and chemical weapons and missiles.

(d) Data security management should be strengthened to comply with data security and personal information protection standards and regulations if training data contains sensitive personal information and important data.

(e) To use truthful, precise, objective, and diverse training data from legitimate sources, and filter ineffective, wrong, and biased data in a timely manner.

(f) The cross-border provision of AI services should comply with the regulations on cross-border data flow. The external provision of AI models and algorithms should comply with export control requirements.

4.1.3 Addressing risks from AI system

(a) To properly disclose the principles, capacities, application scenarios, and safety risks of AI technologies and products, to clearly label outputs, and to constantly make AI systems more transparent.

(b) To enhance the risk identification, detection, and mitigation of platforms where multiple AI models or systems congregate, so as to prevent malicious acts or attacks and invasions that target the platforms from impacting the AI models or systems they support.

(c) To strengthen the capacity of constructing, managing, and operating AI computing platforms and AI system services safely, with an aim to ensure uninterrupted infrastructure operation and service provision.

(d) To fully consider the supply chain security of the chips, software, tools, computing infrastructure, and data sources adopted for AI systems. To track the vulnerabilities and flaws of both software and hardware products and make timely repair and reinforcement to ensure system security.

4.2 Addressing safety risks in AI applications

4.2.1Addressing cyberspace risks

(a) A security protection mechanism should be established to prevent model from being interfered and tampered during operation to ensure reliable outputs.

(b) A data safeguard should be set up to make sure that AI systems comply with applicable laws and regulations when outputting sensitive personal information and important data.

4.2.2 Addressing real-world risks

(a) To establish service limitations according to users' actual application scenarios and cut AI systems' features that might be abused. AI systems should not provide services that go beyond the preset scope.

(b) To improve the ability to trace the end use of AI systems to prevent high-risk application scenarios such as manufacturing of weapons of mass destruction, like nuclear, biological, chemical weapons and missiles.

4.2.3 Addressing cognitive risks

(a) To identify unexpected, untruthful, and inaccurate outputs via technological means, and regulate them in accordance with laws and regulations.

(b) Strict measures should be taken to prevent abuse of AI systems that collect, connect, gather, analyze, and dig into users' inquiries to profile their identity, preference, and personal mindset.

(c) To intensify R&D of AI-generated content (AIGC) testing technologies, aiming to better prevent, detect, and navigate the cognitive warfare.

4.2.4 Addressing ethical risks

(a) Training data should be filtered and outputs should be verified during algorithm design, model training and optimization, service provision and other processes, in an effort to prevent discrimination based on ethnicities, beliefs, nationalities, region, gender, age, occupation and health factors, among others.

(b) AI systems applied in key sectors, such as government departments, critical information infrastructure, and areas directly affecting public safety and people's health and safety, should be equipped with high-efficient emergency management and control measures.

5. Comprehensive governance measures

While adopting technological controls, we should formulate and refine comprehensive AI safety and security risk governance mechanisms and regulations that engage multi-stakeholder participation, including technology R&D institutions, service providers, users, government authorities, industry associations, and social organizations.

5.1 To implement a tiered and category-based management for AI application. We should classify and grade AI systems based on their features, functions, and application scenarios, and set up a testing and assessment system based on AI risk levels. We should bolster end-use management of AI, and impose requirements on the adoption of AI technologies by specific users and in specific scenarios, thereby preventing AI system abuse. We should register AI systems whose computing and reasoning capacities have reached a certain threshold or those are applied in specific industries and sectors, and demand that such systems possess the safety protection capacity throughout the life cycle including design, R&D, testing, deployment, utilization, and maintenance.

5.2 To develop a traceability management system for AI services. We should use digital certificates to label the AI systems serving the public. We should formulate and introduce standards and regulations on AI output labeling, and clarify requirements for explicit and implicit labels throughout key stages including creation sources, transmission paths, and distribution channels, with a view to enable users to identify and judge information sources and credibility.

5.3 To improve AI data security and personal information protection regulations. We should explicate the requirements for data security and personal information protection in various stages such as AI training, labeling, utilization, and output based on the features of AI technologies and applications.

5.4 To create a responsible AI R&D and application system. We should propose pragmatic instructions and best practices to uphold the people-centered approach and adhere to the principle of developing AI for good in AI R&D and application, and continuously align AI's design, R&D, and application processes with such values and ethics. We should explore the copyright protection, development and utilization systems that adapt to the AI era and continuously advance the construction of high-quality foundational corpora and datasets to provide premium resources for the safe development of AI. We should establish AI-related ethical review standards, norms, and guidelines to improve the ethical review system.

5.5 To strengthen AI supply chain security. We should promote knowledge sharing in AI, make AI technologies available to the public under open-source terms, and jointly develop AI chips, frameworks, and software. We should guide the industry to build an open ecosystem, enhance the diversity of supply chain sources, and ensure the security and stability of the AI supply chain.

5.6 To advance research on AI explainability. We should organize and conduct research on the transparency, trustworthiness, and error-correction mechanism in AI decision-making from the perspectives of machine learning theory, training methods and human-computer interaction. Continuous efforts should be made to enhance the explainability and predictability of AI to prevent malicious consequences resulting from unintended decisions made by AI systems.

5.7 To share information, and emergency response of AI safety risks and threats. We should continuously track and analyze security vulnerabilities, defects, risks, threats, and safety incidents related to AI technologies, software and hardware products, services, and other aspects. We should coordinate with relevant developers and service providers to establish a reporting and sharing information mechanism on risks and threats. We should establish an emergency response mechanism for AI safety and security incidents, formulate emergency plans, conduct emergency drills, and handle AI safety hazards, AI security threats, and events timely, rapidly, and effectively.

5.8 To enhance the training of AI safety talents. We should promote the development of AI safety education in parallel with AI discipline. We should leverage schools and research institutions to strengthen talent cultivation in the fields of design, development, and governance for AI safety. Support should be given to cultivating top AI safety talent in the cutting-edge and foundational fields, and also expanding such talent pool in areas such as autonomous driving, intelligent healthcare, brain-inspired intelligence and brain-computer interface.

5.9 To establish and improve the mechanisms for AI safety education, industry self-regulation, and social supervision. We should strengthen education and training on the safe and proper use of AI among government, enterprises, and public service units. We should step up the promotion of knowledge related to AI risks and their prevention and response measures in order to increase public awareness of AI safety in all respects. We should guide and support industry associations in the fields of cybersecurity and AI to enhance industry self-regulation, and formulate self-regulation conventions that exceed regulatory requirements and serve exemplary roles. We should guide and encourage AI technology R&D institutions and service providers to continue to improve their safety capacity. A mechanism for handling public complaints and reports on AI risks and hazards should be established, forming an effective social supervision atmosphere for AI safety.

5.10 To promote international exchange and cooperation on AI safety governance. We should actively make efforts to conduct cooperation with countries, support the building of an international institution on AI governance within the United Nations framework to coordinate major issues related to AI development, safety, security, and governance. We should advance cooperation on AI safety governance under multilateral mechanisms such as APEC, G20 and BRICS, and strengthen cooperation with Belt and Road partner countries and Global South countries. Efforts should be made to study the matters relating to the construction of an AI safety governance alliance to increase the representation and voice of developing countries in global AI governance. AI enterprises and institutions should be encouraged to engage in international exchanges and cooperation, share their best practices, jointly develop international standards of AI safety.

6. Safety guidelines for AI development and application

6.1 Safety guidelines for model algorithm developers

(a) Developers should uphold a people-centered approach, adhere to the principle of AI for good, and follow science and technology ethics in key stages such as requirement analysis, project initiation, model design and development, and training data selection and use, by taking measures such as internal discussions, organizing expert evaluations, conducting technological ethical reviews, listening to public opinions, communicating and exchanging ideas with potential target audience, and strengthening employee safety education and training.

(b) Developers should strengthening data security and personal information protection, respect intellectual property and copyright, and ensure that data sources are clear and acquisition methods are compliant. Developers should establish a comprehensive data security management procedure, ensuring data security and quality as well as compliant use, to prevent risks such as data leakage, loss, and diffusion, and properly handle user data when terminating AI products.

(c) Developers should guarantee the security of training environment for AI model algorithms, including cybersecurity configurations and data encryption measures.

(d) Developers should assess potential biases in AI models and algorithms, improve sampling and testing for training data content and quality, and come up with effective and reliable alignment algorithms to ensure risks like value and ethical risks are controllable.

(e) Developers should evaluate the readiness of AI products and services based on the legal and risk management requirements of the target markets.

(f) Developers should effectively manage different versions of AI products and related datasets. Commercial versions should be capable of reverting to previous versions if necessary.

(g) Developers should regularly conduct safety and security evaluation tests. Before testing, they should define test objectives, scope, safety and security dimensions, and construct diverse test datasets covering all kinds of application scenarios.

(h) Developers should formulate clear test rules and methods, including manual testing, automated testing, and hybrid testing, and utilize technologies such as sandbox simulations to fully test and verify models.

(i) Developers should evaluate tolerance of AI products and services for external interferences and notify service providers and users in forms of application scope, precautions, and usage prohibitions.

(j) Developers should generate detailed test reports to analyze safety and security issues, and propose improvement plans.

6.2 Safety guidelines for AI service providers

(a) Service providers should publicize capabilities, limitations, target users, and use cases of AI products and services.

(b) Service providers should inform users of the application scope, precautions, and usage prohibitions of AI products and services in a user-friendly manner within contracts or service agreements, supporting informed choices and cautious use by users.

(c) Service providers should support users to undertake responsibilities of supervision and control within documents such as consent forms and service agreements.

(d) Service providers should ensure that users understand AI products' accuracy, and prepare explanatory plans when AI decisions exert significant impact.

(e) Service providers should review responsibility statements provided by developers to ensure that the chain of responsibility can be traced back to any recursively employed AI models.

(f) Service providers should increase awareness of AI risk prevention, establish and improve a real-time risk monitoring and management mechanism, and continuously track operational security risks.

(g) Service providers should assess the ability of AI products and services to withstand or overcome adverse conditions under faults, attacks, or other anomalies, and prevent unexpected results and behavioral errors, ensuring that a minimum level of effective functionality is maintained.

(h) Service providers should promptly report safety and security incidents and vulnerabilities detected in AI system operations to competent authorities.

(i) Service providers should stipulate in contracts or service agreements that they have the right to take corrective measures or terminate services early upon detecting misuse and abuse not conforming to usage intention and stated limitations.

(j) Service providers should assess the impact of AI products on users, preventing harm to users' mental and physical health, life, and property.

6.3 Safety guidelines for users in key areas

(a) For users in key sectors such as government departments, critical information infrastructure, and areas directly affecting public safety and people's health and safety, they should prudently assess the long-term and potential impacts of applying AI technology in the target application scenarios and conduct risk assessments and grading to avoid technology abuse.

(b) Users should regularly perform system audits on the applicable scenarios, safety, reliability, and controllability of AI systems, while enhancing awareness of risk prevention and response capabilities.

(c) Users should fully understand its data processing and privacy protection measures before using an AI product.

(d) Users should use high-security passwords and enable multi-factor authentication mechanisms to enhance account security.

(e) Users should enhance their capabilities in areas such as network security and supply chain security to reduce the risk of AI systems being attacked and important data being stolen or leaked, as well as ensure uninterrupted business.

(f) Users should properly limit data access, develop data backup and recovery plans, and regularly check data processing flow.

(g) Users should ensure that operations comply with confidentiality provisions and use encryption technology and other protective measures when processing sensitive data.

(h) Users should effectively supervise the behavior and impact of AI, and ensure that AI products and services operate under human authorization and remain subject to human control.

(i) Users should avoid complete reliance on AI for decision making, monitor and record instances where users turn down AI decisions, and analyze inconsistencies in decision-making. They should have the capability to swiftly shift to human-based or traditional methods in the event of an accident.

6.4 Safety guidelines for general users

(a) Users should raise their awareness of the potential safety risks associated with AI products, and select AI products from reputable providers.

(b) Before using an AI product, users should carefully review the contract or service terms to understand its functions, limitations, and privacy policies. Users should accurately recognize the limitations of AI products in making judgments and decisions, and set reasonable expectations.

(c) Users should enhance awareness of personal information protection and avoid entering sensitive information unnecessarily.

(d) Users should be informed about data processing practices and avoid using products that are not in conformity with privacy principles.

(e) Users should be mindful of cybersecurity risks when using AI products to prevent them from becoming targets of cyberattacks.

(f) Users should be aware of the potential impact of AI products on minors and take steps to prevent addiction and excessive use.

Table of AI Safety and Security Risks to Technical Countermeasures and Comprehensive Governance Measures

Safety risks			Technical countermeasures	Comprehensive governance measures
Inherent safety risks	Risks from models and algorithms	Risks of explainability	4.1.1 (a)	● Advance research on AI explainability ● Create a responsible AI R&D and application system
		Risks of bias and discrimination	4.1.1 (b)	
		Risks of robustness	4.1.1 (b)	
		Risks of stealing and tampering	4.1.1 (b)	
		Risks of unreliable output	4.1.1 (a) (b)	
		Risks of adversarial attack	4.1.1 (b)	
	Risks from data	Risks of illegal collection and use of data	4.1.2 (a)	● Improve AI data security and personal information protection regulations
		Risks of improper content and poisoning in training data	4.1.2 (b) (c) (d) (e) (f)	
		Risks of unregulated training data annotation	4.1.2 (e)	
		Risks of data leakage	4.1.2 (c) (d)	
	Risks from AI systems	Risks of exploitation through defects and backdoors	4.1.3 (a) (b)	● Strengthen AI supply chain security ● Share information, and emergency response of AI safety risks and threats
		Risks of computing infrastructure security	4.1.3 (c)	
		Risks of supply chain security	4.1.3 (d)	
Safety risks in AI applications	Cyberspace risks	Risks of information and content safety	4.2.1 (a)	● Implement a tiered and category-based management system for AI application ● Establish a traceable management system for AI services ● Increase efforts to train talent in AI safety and security ● Establish and improve mechanisms for AI safety and security education, industry self-regulation, and social supervision ● Promote international exchange and cooperation on AI safety governance
		Risks of confusing facts, misleading users and bypassing authentication	4.2.1 (a)	
		Risks of information leakage due to improper usage	4.2.1 (b)	
		Risks of abuse for cyberattacks	4.2.1 (a)	
		Risks of security flaw transmission caused by model reuse	4.2.1 (a) (b)	
	Real-world risks	Inducing traditional economic and social security risks	4.2.2 (b)	
		Risks of using AI in illegal and criminal activities	4.2.2 (a) (b)	
		Risks of misuse of dual-use items and technologies	4.2.2 (a) (b)	
	Cognitive risks	Risks of amplifying the effects of "information cocoons"	4.2.3 (b)	
		Risks of usage in launching cognitive warfare	4.2.3 (a) (b) (c)	
	Ethical risks	Risks of exacerbating social discrimination and prejudice, and widening the intelligence divide	4.2.4 (a)	
		Risks of challenging traditional social order	4.2.4 (a) (b)	
		Risks of AI becoming uncontrollable in the future	4.2.4 (b)	